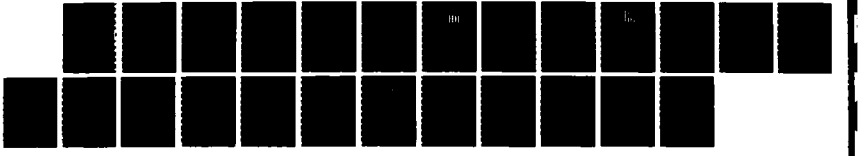
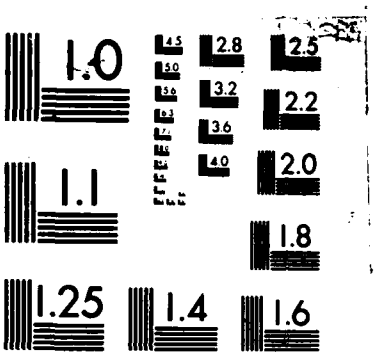
AD-A177 229 ON-CHIP INSTRUMENT CACHES FOR HIGH PERFORMANCE PROCESSORS(U) STANFORD UNIV CA COMPUTER SYSTEMS LAB A AGARWAL ET AL. 1987 MDA903-83-C-0335 1/1

UNCLASSIFIED F/G 9/5 NL





Contract MDA-903-83-C-0335

On-chip Instruction Caches for High Performance Processors

Anant Agarwal, Paul Chow, Mark Horowitz,
John Acken, Arturo Salz, and John Hennessy

Computer Systems Laboratory
Stanford University
Stanford, CA 94305

1 Abstract

Continued increases in clock rates of VLSI processors demand a reduction in the frequency of expensive off-chip memory references. Without such a reduction, the chip crossing time and the constraints of external logic will severely impact the clock cycle. By absorbing a large fraction of instruction references, on-chip caches substantially reduce off-chip communication. Minimizing the average instruction access time with a limited silicon budget requires careful analysis of both cache architecture and implementation. This paper examines some important design issues and tradeoffs that maximize the performance of on-chip instruction caches, while retaining implementation ease. Our discussion focuses on the instruction cache design for MIPS-X, a pipelined, 32-bit, reduced instruction set, 20 MIPS peak, CMOS processor designed at Stanford. The on-chip instruction cache is 2K bytes and allows single-cycle instruction accesses. Trace driven simulations show that the cache has an average miss rate of 12% resulting in an average instruction access time of 1.24 cycles.

2 Introduction

With the rapid improvement in processor architecture, led by the RISC processors, and with advances in VLSI technology, the cost of off-chip communication has not kept pace with improvements in the clock rates of VLSI processors. Consequently, the performance of current high-performance VLSI processors is memory bandwidth limited. Including memory on the processor chip to reduce the cost of memory accesses becomes imperative to attain higher performance [1,2,3].

The RISC approach [4,5,1] advocates shifting functionality to software and using hardware resources to support features that clearly improve performance. An on-chip instruction cache is just such a feature. Furthermore, by achieving reasonably high hit rates in an on-chip instruction cache, we can eliminate the major disadvantage of a RISC architecture: decreased instruction density. As we show, the silicon freed up by the lack of microcode more than pays for itself in the amount that it compresses the off-chip instruction density. Several current VLSI processor designs have observed these facts and employed on-chip instruction caches, including the MIPS-X project at Stanford [6], the Berkeley SPUR project [7], the Motorola 68020 [8] and the 68030 [9].

With current technology, small instruction-only caches are attractive for VLSI implementation for a number of reasons:

1. An on-chip instruction cache provides an extra port to memory and allows simultaneous fetches of both instructions and data with a single set of external address and

data pins. The I/O pins do not see most of the instruction fetches. The importance of dual porting becomes apparent from peak bandwidth requirements. MIPS-X [6] requires a peak of 40 million words per second, 20 million each for instructions and data. With a minimum cycle time of 50 nanoseconds, the off-chip bandwidth would be inadequate without using a large number of pins. A combined on-chip instruction-data cache would also be insufficient, unless a much larger dual-ported cache could be built.

2. Streamlined machines require a high instruction bandwidth that can be over twice the data bandwidth. For example, in MIPS-X, instructions comprise 70% of all memory references, and in the Berkeley SPUR processor, which includes a large register file, instructions comprise 82% of all memory references for C benchmarks and 75% for Lisp benchmarks [10]. Given the 50ns cycle time, and the current CMOS technology, only an on-chip cache can sustain the required instruction bandwidth.

3. Because instructions show a much higher degree of locality than data, especially for well structured programs, small instruction caches yield lower miss rates than combined instruction-data caches, and considering the net performance gain achievable, the limited silicon resources are best exploited by an instruction cache. Achieving a high hit rate in the on-chip cache is even more important when an off-chip cache is present (as in MIPS-X), and the miss penalty is small enough that a small on-chip cache with a high miss rate can actually be detrimental to performance due to overhead in cache management.

4. Instruction caches are simpler to implement than a combined instruction-data cache because the the cache need not support writes. The architecture can disallow writes into the instruction stream. In addition, a read-only instruction cache does not have cache consistency problems for multiprocessor configurations.

This paper examines the design of integrated instruction caches from a practical standpoint. The goal is to minimize the average cost of an instruction fetch. This cost is a function of the cache hit rate, the miss penalty, and the cache access time. A number of VLSI specific implementation issues, such as available area, cache aspect ratio, timing, and ease of implementation, must also be addressed. Because many of these constraints are specific to the design of a given processor, much of our discussion is in the context of the MIPS-X instruction cache. However, the methodology, and the trade-offs and optimizations that we discuss are relevant to other designs as well.

The next section reviews cache performance evaluation methodology and terminology. Section 4 analyzes various cache organizations to assess their performance and see how suited they are to VLSI implementation. Then, following a discussion of MIPS-X cache miss timing, we look at a number of schemes (including prefetching) to improve the average access time of an instruction. In section 5, we examine several replacement algorithms, handling context switches, and discuss the tradeoffs and optimizations made.

3 Methodology

3.1 A Cache Performance Metric

A cache design involves optimizing cache performance with respect to a number of cache parameters. Cache performance is usually based on the cache miss rate, or the fraction of

cache requests not satisfied by the cache, and traffic ratio, or the ratio of memory references made by the cache to the number of cache accesses made by the processor. A more important metric is the *average memory access time* because some cache organizations achieve a lower miss rate at the cost of increasing the miss service time. Moreover, both the miss rate and the traffic ratio do not consider implementation. This omission is significant because, as we show later, performance is often more sensitive to the implementation than to the cache architecture.

In this paper we will concentrate predominantly on the average instruction access time which depends both on the miss rate and the miss penalty. If the cache miss rate is m, and the miss service time is T_{miss} cycles, then average instruction access time, T_{ave}, is $1 + mT_{miss}$ cycles; we assume an instruction access takes one cycle if it is present in the instruction cache.

3.2 Cache Design Parameters

The cache parameters that are of interest in instruction cache design include the cache size, set size, block size, sub-block size, and replacement algorithm [11]. Note that write policies are not relevant to us because we disallow writes into the instruction stream. The number of sets has also been called the number of rows; the set size is the scope of associative search, and has also been called degree of associativity or number of columns. Block size is defined as the amount of storage associated with an address tag and is sometimes referred to as line size. Sub-block placement is also called sector placement [12]. A sub-block (or transfer block [3]) is the portion of a block transferred from memory on a cache miss. Since a block can simultaneously have invalid sub blocks in addition to valid ones, each sub block must have a valid bit associated with it. The replacement algorithm is the process used to select one of the blocks in a given set for occupation by a newly referenced block. The important schemes include LRU (least recently used), random, and FIFO (first in first out).

Cache size is limited by the amount of chip space available for both storage of instructions and the address tags. The choice of set size, block size, and sub-block size depends on a number of factors including (1) the miss rate achievable, (2) the timing of a cache access and how it fits in with the timing of the rest of the machine, and (3) implementation ease.

3.3 Evaluation methodology

Initially, we investigate the miss rates of various cache organizations. Then, from an implementation standpoint, we analyze the cache access timing and the miss penalty associated with each organization. Finally, we determine the average instruction access time from the miss rate and the cost of servicing a miss.

Trace driven simulation (TDS) is used to obtain the cache miss rates. Because of its flexibility and ease of use, TDS is a popular technique for cache performance evaluation [13,14]; however, TDS does have some drawbacks. It may not be as accurate as hardware measurement because traces seldom reflect true workload behavior. TDS results are often optimistic because large applications, usually with poor cache performance, are hard to trace; moreover, the effect of multiprogramming, another cause of cache performance degradation, is hard to include in TDS studies.

Fortunately, these problems are not very serious in studying small instruction caches. Since small caches self purge after a few thousand references, multiprogramming has little

effect on performance. For example, if the miss rate of a 512-word cache is 20% without multiprogramming, and if the cache is flushed every 20 thousand references, the worst case miss rate will now be 22.5%, which is a degradation of less than 15%. Even simple models for multiprogramming, such as starting with an empty cache every few thousand references, are sufficient. In some cases (e.g., MIPS-X) this simple model can be an accurate representation of an actual virtual address cache, where the cache is flushed every time a new user process is started. Our initial analyses ignore the effects of multiprogramming. Later, to study the impact of context switching, we flush the cache every few thousand references.

In the initial phases of the MIPS-X processor design we did not have either a running software system or an instruction simulator for MIPS-X. Much of the design was based on traces obtained using the MIPS system (MIPS [15] is the predecessor to MIPS-X) assuming a similar behavioral trend for MIPS-X. The MIPS-X software system and a simulator have since been developed and we have corroborated our earlier findings, with the only difference being that the performance of the MIPS-X cache turned out to be slightly worse than predicted because MIPS-X code is less dense, and our current benchmarks are larger.

In this paper, we present the results using large MIPS-X benchmark traces obtained from a MIPS-X instruction level simulator. The simulator takes an object module and generates an address trace by interpretively executing the program. Each address is tagged as an instruction read, data read, or data write. Ten Pascal and Lisp benchmarks are used:

Bigfm	Fiduccia-Mattheyes graph partitioning algorithm (Pascal).
Dnf	Converts logic equations to disjunctive normal form (Pascal).
Hopt	A simple global optimizer for Pascal (Pascal).
Simu	Operating system paging simulator (Pascal).
Upas	Pascal compiler front end (Pascal).
Comp	First pass of the PSL compiler (macro expand) (Lisp).
Frl	Frame representation language (Lisp).
Gc	Deduction with garbage collection (Lisp).
Rat	Rational expression evaluator (Lisp).
Opt	Compiler - data flow and optimization (Lisp).

These programs are fairly large, ranging from 50,000 to 800,000 bytes in static code size, and we feel they provide realistic cache performance numbers.

4 Cache Organization

This section discusses tradeoffs in the selection of cache parameters including number of sets, set size, and block size. The basic question is how to best utilize a given amount of chip area of a particular aspect ratio to obtain the maximum performance in terms of average instruction access time. Early cache studies, including those of Kaplan and Winder [16], Strecker [17], and Smith [11], compared cache performance for various cache organizations assuming a fixed amount of data storage. Alpert [18] stressed that for integrated microprocessor caches the total circuit area associated with the cache, including both tags and data, must be considered. Our experience shows that in addition to area, the aspect ratio is also important.

To reasonably limit the number of variables, we explore the design space available for the MIPS-X processor. We encourage the reader to concentrate on the evaluation procedure rather than on the final result, which may well be different given other basic constraints.

Although the total cache size is fixed, a wide variety of organizations exploiting various features of program behavior are possible. These alternatives, specifically relevant to VLSI processor caches, must be analyzed for their performance, feasibility, and ease of implementation. We also present possible floor plans to make the most effective use of available area. We will consider in turn a conventional cache or a c-cache, a buffer, and what we call a hybrid buffer.

A *c-cache*, for the purpose of this analysis, is an organization that uses about half the available memory space for the tags. Each block consists of one to four words (each word is 4 bytes). A possible 512-word c-cache organization could have 512 sets (rows), set size or associativity one, and block size one word. A portion of the address first indexes into a cache set, followed by an associative tag compare against the blocks in that set.

A *buffer* is a set of a few large blocks, block size being eight words or more. For example, a 512-word buffer could be organized with set size eight, and block size 64 words. Since a buffer has only one set, the associative search can be started without the indexing operation. This organization has the minimum number of tags.

A *hybrid buffer* is also investigated because the more straightforward organizations typically used in instruction buffers (e.g., CRAY-1 [19],VAX-11/780 [20]), or in previous VLSI processor instruction caches (e.g., Motorola MC68020 [8], Zilog Z80,000 [18]), are not optimal in our case. As the name suggests, a hybrid buffer has the features of both a c-cache and a buffer. Like a buffer it has a large block size and few tags; consequently the tag store is small. It is similar to a c-cache and differs from a buffer because it has more than one set, typically two or four. A typical 512-word hybrid buffer could have two sets, set size 16, and block size 16 words.

We assume a sub-block size of one word. In other words, a single instruction is fetched into the cache on a miss.[1] Each word in the cache or buffer is associated with a valid bit. The replacement algorithm (discussed in detail later) is pseudo-random.

4.1 Technology Constraints

The technology, organization, and performance of MIPS-X greatly constrained the cache design. The most important constraint was size. Roughly half of the interior die area was allocated to the cache, giving it a space of 4mm by 6mm. The floorplan of the processor fixed the aspect ratio. The datapath was expected to be at least 6mm long and the cache had to fit above the datapath. In our design rules, the area of a 6 transistor static memory cell was 30μm by 40μm, which allowed us to build roughly a 16K-bit memory in the available area. We considered using dynamic memory cells, but decided it was not worth the technology risk.

From preliminary design and layouts for the sense-amplifiers, column multiplexers and tag compare logic associated with the cache, the column multiplexers and the sense-amplifiers were estimated to require around 300μm of space below the cache RAM array, and the tag logic an additional 300μm. The large space for the tag logic was a result of the 24 wires needed to provide the comparison address for the tags. The wire pitch was about 10μm (second-level metal).

In addition to the area constraints, MIPS-X also constrained the access time of the cache; an instruction fetch had to complete in a single 50ns clock cycle. To meet the cycle time constraint, we felt the cache would only have time for a single RAM access per cache access. Thus, we were concerned about implementing a set-associative cache since

[1] Many previous papers support sub-block placement in small on-chip caches [18,3,21].

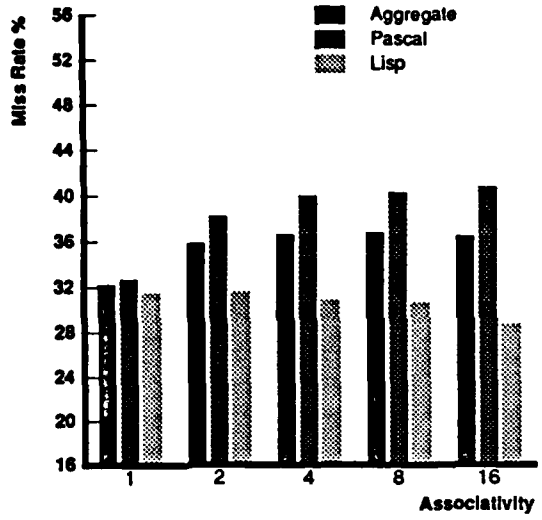


Figure 1: Miss rate of a 2K-byte instruction c-cache with block size = 4 bytes. The aggregate miss rates are calculated as the average miss rate of all benchmarks weighted by the number of references.

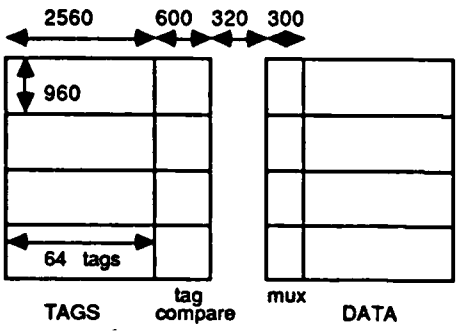
we would need to first fetch the tags and then fetch the correct data word. To alleviate the need for sequential fetches, we organized the cache so that the tag information could arrive late in the cycle. The tags were only used for the column decode. In this organization, when the word-line rose, all possible data words were fetched onto the bit-lines. The tag information was used to select the correct set of bit-lines to the sense amplifiers. The delay from the tag becoming valid to output valid was short since it bypassed the delay incurred in driving the bit-lines. The limitation of this technique was that it required 32 bit-lines for each degree of associativity used. A 4-way set-associative cache would need 128 pairs of bit-lines.

The cache designs described below attempt to meet both the size and organizational constraints described in this section.

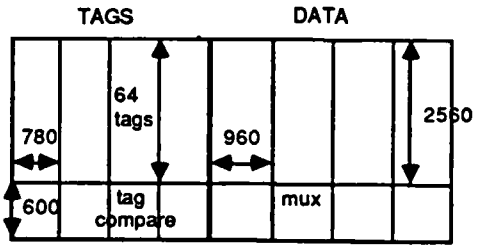
4.2 C-cache

For the c-cache study we initially use a block size of one word (4 bytes). Therefore, half the area is taken up by the tags; the data store consists of 256 words. To reduce the area required by the tag store, we also try caches with a larger cache block size. We assume that if the block size is greater than two words, we can squeeze in 512 words for instructions.

Figure 1 shows the miss rate for an instruction c-cache with associativity ranging from one through 16. The number of sets correspondingly range from 256 to 16. The best possible aggregate miss rate of 32.26%, interestingly enough, is achieved by a direct mapped c-cache. We felt that some anomaly in the replacement scheme (pseudo-random) caused the miss rate to go up with associativity for the small cache. A later simulation with LRU replacement also showed this behavior for all the Pascal benchmarks with the exception of Upas. This behavior can be explained by examining the reference and collision pattern in a small cache, and is further discussed in [22,23]. Lisp benchmarks do not show this anomalous behavior to the same extent as Pascal, because Lisp tends to have a higher frequency of procedure calls and shorter bodies of sequential code. This causes an increased probability of interference that can be reduced by associativity.



(a) Size: 6040 by 3840 μm.



(b) Size: 6960 by 3160 μm.

Figure 2: Possible floorplans of a c-cache with 1K bytes of data store. Number of sets is 64, associativity is four, and block size is four bytes; all dimensions are in μm.

A c-cache seems more effective for Lisp benchmarks than for Pascal. As we will see, the converse is true for a buffer. The reason is that a buffer exploits code sequentiality, while a c-cache is preferable when many short non-sequential localities of code are present, as in Lisp.

Possible floorplans for a four-way set-associative c-cache are shown in Figure 2. Note that lesser associativity caches can use the same basic floorplan; there is no reason to change the organization. For the floorplan in Figure 2(a), the dimensions are 3480 by 6040 μm and for the floorplan in Figure 2(b), the dimensions are 3160 by 6960 μm. The actual dimensions are slightly greater when the precharge and decode logic is added. Although both arrangements have the same area, only the former was suitable to our purpose due to the aspect ratio constraint.

We have assumed, thus far, that half the area is occupied by tags. Other c-cache types with larger block sizes and a fewer number of tags are also possible. For instance, a c-cache with a block size of four would have the tags occupying only a fourth as much area as the data portion. For these large block sizes we thought that we might be able to squeeze in 2K bytes of data (the same as for a buffer scheme) with the corresponding tags.

The miss rates for block sizes of 4, 16, and 32 bytes are given in Figure 3, and a floorplan in Figure 4. While the area occupied is much larger for this configuration because a full 2K bytes of instructions are stored, the miss rate is reduced substantially (from 32.26%

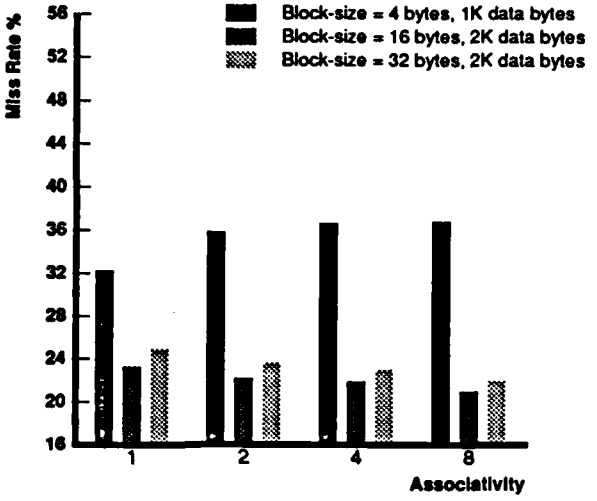


Figure 3: Effect of large cache-block size.

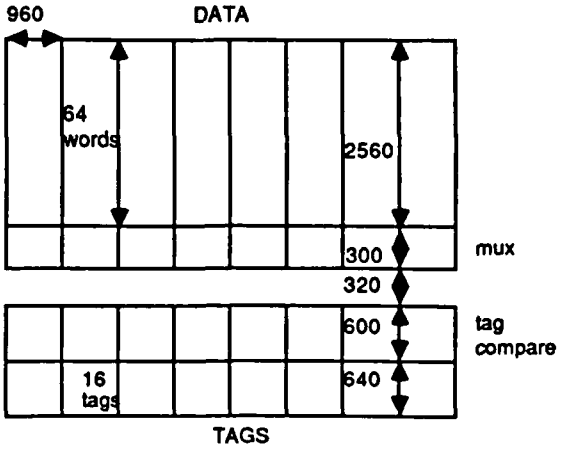


Figure 4: Floorplan of a c-cache with 2K bytes of data store, 16 sets, associativity eight, and block size 16 bytes. Size: 7680 by 4420 μm.

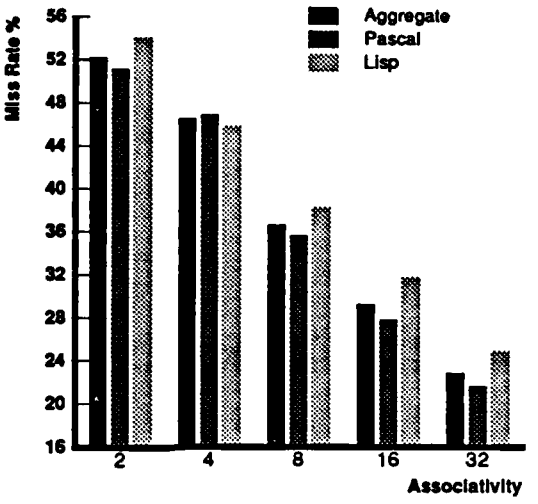


Figure 5: Miss rate of a 2K-byte instruction buffer.

to 23.28% for a direct mapped cache). Clearly, for caches as small as 2K bytes, size is the single most important parameter affecting the miss rate. Other parameters play only a secondary role; e.g., doubling the associativity for the 2K byte cache, with block size 16 bytes, causes the miss rate to decrease slightly from 23.28% to 22.16%; and doubling the block size causes the miss rate to marginally increase from 23.28% to 24.91%. Some parameters, however, affect the cache timing. For instance, an associative cache requires tag access, compare, a select and a drive; while a direct mapped cache requires just the access, compare, and drive. A c-cache with sixteen sets, degree of associativity eight, and block size sixteen bytes achieves the lowest miss rate of 20.96%. The combined area occupied by the tags and data is 4420 by 7680 μm, which exceeds the space we had available.

4.3 Buffer

Alpert [18] showed that reducing the number of tags is desirable when the area available for the cache is small and fixed. Our experiments show that this is true even to a greater extent for instruction caches. Instructions tend to show high spatial locality, which large block sizes can effectively exploit. Large block sizes are particularly attractive for reduced instruction set computers because of poorer code density and the correspondingly larger basic block sizes in the code. Buffers are attractive because they minimize the number of tags, and have the added advantage that they are well suited to prefetch schemes such as load forward,[2] as well as the MIPS-X prefetch scheme outlined later.

Because a buffer needs few tags, most of the area is used for storing instructions. Hence, we do not show tags in our buffer floorplans. Access time can be made lower by (1) eliminating the indexing operation to choose a set, and (2) decreasing the tag access time by using fast circuits[3] and placing the few tags and the valid bits close to the address generation logic. For example, in MIPS-X the tags are located in the datapath itself.

We studied instruction buffers ranging from a block size of 1024 bytes and associativity

[2]The load forward scheme was implemented in [24] and also studied by Hill and Smith [21].

[3]Although faster circuits are larger and consume more power, the overall size and power increase is small because of fewer tags.

two to a block size of 64 bytes and associativity 32. Figure 5 shows the miss rates. The most striking feature of the graph is the importance of block size. Smaller block sizes allow a larger associativity. Clearly, the miss rate can be very high for large block sizes. The reason is that for low associativity and a correspondingly large block size, major portions of blocks tend to be left unused. The lowest miss rate is 22.79% for a 32-way set-associative buffer, which is substantially lower than that for a direct mapped c-cache, and very similar to the best miss rate achieved for a c-cache (20.96%). It achieves this miss rate at a smaller silicon area than the comparable c-cache because it uses a smaller number of tags.

As mentioned before, a buffer is more effective for Pascal benchmarks than for Lisp benchmarks, while the opposite is true for a c-cache organization. Because Lisp code has on average shorter bodies of sequential code than Pascal, large buffer blocks tend to be under-utilized resulting in poorer cache performance for Lisp. Table 1 provides empirical evidence of this behavior by showing the average number of words utilized per block in a fully-associative 2K-byte buffer with block size 16 words for both Pascal and Lisp. The block utilization by Pascal benchmarks is roughly 25% more on average than the Lisp benchmarks.

Pascal Benchmark	Words Utilized	Lisp Benchmark	Words Utilized
Bigfm	12.2	Comp	8.4
Dnf	11.3	Frl	8.9
Hopt	10.7	Gc	9.4
Simu	10.1	Opt	8.5
Upas	9.8	Rat	8.8
Average	10.9	Average	8.8

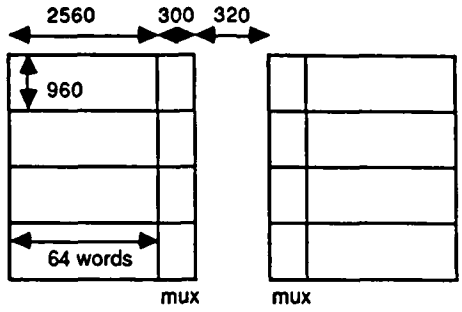
Table 1: Average number of words of Pascal and Lisp benchmarks utilized per block in a 2K-byte buffer with block size 16 words.

Possible floorplans for the various buffers are given in Figure 6. A set size of eight is implementable using the layout shown in Figure 6(a) with dimensions 3840 by 6040 μm. The layout shown in Figure 6(b), has dimensions 3180 by 7680 μm, and is too long in one dimension. Unfortunately, larger set sizes required to achieve reasonable performance are hard to implement. Figure 7 shows the floorplan for an associativity of 16. The layout is wider because of the extra bus routing channels required. When the area required by the decode and precharge logic for each of the banks is added, the dimensions along the width become much larger than we can allow.

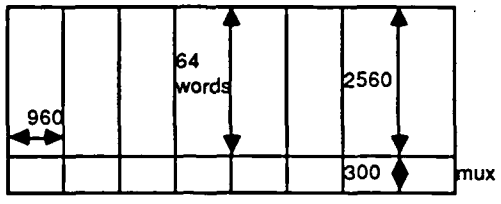
4.4 Hybrid Buffer

A c-cache suffers from the drawbacks that tags occupy valuable space and tag access requires a RAM access. A buffer reduces these problems by reducing the number of tags and using special structures to reduce the effective tag access time. A pure buffer requires a high degree of associativity, making the actual RAM harder to design (it needs to have a large number of bit lines). Since the large associativity is required only to keep the block size down, we will provide the same block size with a lower associativity in a structure called a hybrid buffer.

A hybrid buffer simulates a higher associativity in the following manner. Consider



(a) Size: 6040 by 3840 μm.



(b) Size: 7680 by 2860 μm.

Figure 6: Possible floorplans of an 8-way set-associative buffer, with block size 256 bytes.

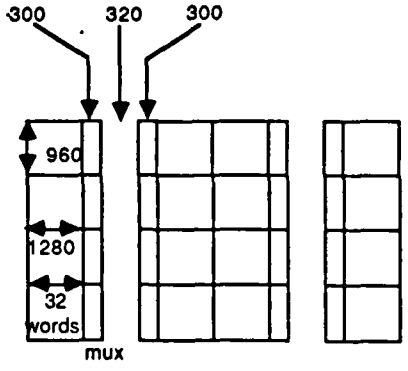


Figure 7: 16-way set-associative buffer with block size 128 bytes. Size: 6960 by 3840 μm.

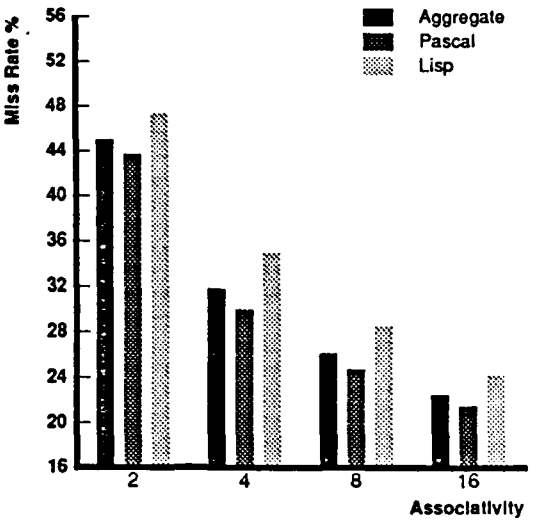


Figure 8: Miss rate of a 2K-byte hybrid buffer with two sets.

two regular buffers where instructions map to either one of the buffers depending on the value of a bit in the address. This is similar to a c-cache with two sets indexed by a bit in the address. If each set (or buffer) has equal probability of being the target of a block, the number of instructions that will fall into any one buffer is halved, which effectively doubles the available associativity. Thus, this scheme provides the benefits of a higher associativity without implementation problems.

Figure 8 shows the miss rate of a hybrid buffer with two sets for various associativities. As expected, the miss rate for a hybrid buffer with two sets and a given associativity, is similar to that for a buffer with twice the associativity. In particular, the miss rate for a hybrid buffer with associativity 16 is 22.48% which is very close to 22.79% for a buffer with associativity 32.

From an implementation point of view, associativity of 16 is still hard to achieve.[4] A hybrid buffer with associativity 8 is implementable; unfortunately the miss rate is about 15% worse (26% vs. 22%). The solution is to extend the number of sets to four with the assumption that the probability distribution of blocks into the four sets is uniform. Two bits are now used to index into one of four sets. The miss rates for a hybrid buffer with four sets are plotted in Figure 9.

In this case the miss rate for a hybrid buffer with degree of associativity eight is slightly worse than for the buffer with associativity 32: 23.17% compared to 22.79%. The difference is because our assumption that each set is equally likely to be the target of a block is not quite true. The variance in the number of instructions that map to any one set causes a higher number of misses.

A floorplan for an eight-way set-associative hybrid buffer is shown in Figure 10. Two bits index into one of four sets, four bits form the block offset, and the rest of the word address forms the tag. The size of 3840 by 6040 μm can comfortably fit in the available silicon area.

[4]Note that at this point we are assuming that the valid bits are in the cache array. In a later section we will show how this associativity can be achieved.

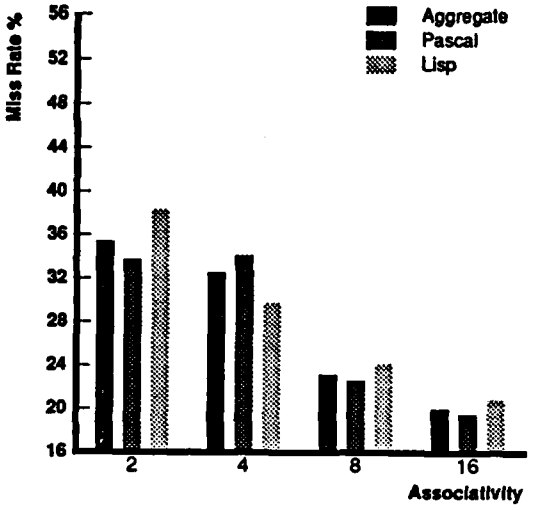


Figure 9: Miss rate of a 2K-byte hybrid buffer with four sets.

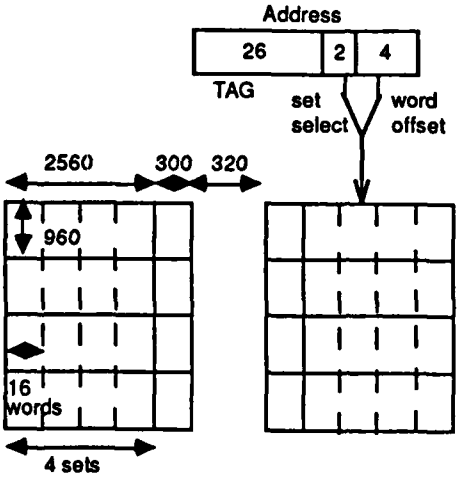


Figure 10: Floorplan of an 8-way set-associative hybrid buffer. Size: 6040 by 3840 μm.

IF_{-1}	ϕ_1	Compute PC
	ϕ_2	Drive PC onto PCBus
IF	ϕ_1	
	ϕ_2	Instruction fetch
RF	ϕ_1	Instruction and register decode
	ϕ_2	Register fetch
ALU	ϕ_1	ALU or shift operation for compute instructions; address computation for load or store instructions; destination and condition computation for branches
	ϕ_2	Drive address to pads for load or store
MEM	ϕ_1	Wait for external cache memory access
	ϕ_2	Drive data to pads for stores; Data arrives at pads late in this phase for loads
WB	ϕ_1	Write result into the destination register
	ϕ_2	

Figure 11: MIPS-X pipeline stages.

4.5 Timing Considerations

Now that we have determined several possible cache organizations based on their hit ratios and on their physical layouts, it is necessary to examine how they fit in with the timing of the rest of the machine. The only timing specification used so far has been that the cache access must be within the 50 ns cycle time of the machine. The other important timing consideration is how cache misses can be handled. To understand how various cache organizations affect the instruction cache miss timing of MIPS-X, we will first describe the MIPS-X pipeline.[5] Then we will show how the timing of the cache hit detection affects the number of cycles needed to service a cache miss. We will also use these timings to show how two instructions can be fetched back on a cache miss to decrease the miss rate by a factor of almost two.

4.5.1 The MIPS-X Pipeline

MIPS-X is heavily pipelined so that one instruction can be issued every 50 ns. Each instruction is divided into five pipeline stages and each stage is divided into two 25 ns phases called ϕ_1 and ϕ_2. The pipeline stages and their functions are described in Figure 11. It is conceptually easier to understand the pipeline if you think of an additional stage called IF_{-1} that occurs before the IF stage. During IF_{-1}, the Program Counter unit generates the address of the instruction to be fetched on the following IF. To denote an inserted cache miss cycle we will use CM_n, where n is the number of the cache miss cycle.

4.5.2 Instruction Cache Miss Timing

Hit detection timing affects the number of cache miss cycles that are needed to service an instruction miss in MIPS-X. The short cycle time, coupled with the necessary chip crossings, means that external memory fetches take longer than one cycle. The address must be presented to the processor pads sufficiently early to ensure it is valid on the external pins by the time ϕ_2 falls (see Figure 11). This means that the datapath must

[5] A detailed discussion is presented in [6,25].

IF_{-1}	ϕ_1	Compute PC
	ϕ_2	Drive PC onto PCBus
IF	ϕ_1	Start tag access
	ϕ_2	Detect miss
CM_1	ϕ_1	
	ϕ_2	Drive PC out to external cache
CM_2	ϕ_1	
	ϕ_2	Instruction back from external cache
CM_3	ϕ_1	
	ϕ_2	Write instruction into c-cache and instruction register
RF	ϕ_1	Miss sequence completed
	ϕ_2	

Figure 12: Miss Timing for a C-cache

drive the address bus early in ϕ_2 to start a memory fetch on the following cycle. The external cache memory then drives the processor pins with the required word by the end of ϕ_2 of the following cycle. Thus, a memory access takes one and a half cycles from the time the address is computed.

For the c-cache configurations, the tags and tag comparison have to be put in the cache array to make the implementation feasible. This means the critical path for miss detection involves driving the address up to the cache array, fetching the tags, doing the comparison, and then getting the hit or miss result back to the datapath. Referring to Figure 12, the instruction address is generated during ϕ_2 of IF_{-1}, driven to the cache array during ϕ_1 of IF, the tag fetch begins late in ϕ_1 of IF, and the comparison is computed and driven to the datapath during ϕ_2 of IF, too late to start an external fetch in the same cycle.

If the number of tags is small, as in the buffer or hybrid buffer schemes, the tags can actually be placed in the datapath close to the PC unit, making hit or miss detection in the tags much quicker. Tag compare is simply achieved by doing an associative compare of the tags to the PC bus and "OR-ing" all the match lines. However, the critical path for miss detection still involves driving the address to the cache array to fetch the valid bits.[6] Therefore, as for the c-cache, the miss signal arrives at the datapath by the end of ϕ_2 of IF.

The miss penalty for the c-cache or buffer schemes with valid bits stored in the cache array is three cycles. The timing of an instruction cache miss for the c-cache is shown in Figure 12. Because the miss signal arrives at the datapath late in ϕ_2 of IF, the PC can be driven out to the external cache only in the next cycle. Three cache miss cycles were inserted at this point: *CM1* to drive the instruction address out to the external cache (under normal circumstances a data address is potentially sent out), *CM2* for the external cache access, and *CM3* to write the instruction into the instruction cache and instruction register.

A three cycle penalty, with a miss rate of over 20% degrades processor performance by over 60%. A cache miss penalty of two cycles, which reduces the loss to 40%, can be achieved by combining *CM2* and *CM3* into one cycle. The critical path now involves accessing the external cache memory, getting the data to the processor pads, and writing into the instruction cache. Since the data from the external cache arrives late in the cycle, this approach can easily affect the cycle time of the machine. Although it decreases the

[6]Note that our sub-block size of one word requires a valid bit for every word.

IF_{-1}	ϕ_1	Compute PC
	ϕ_2	Drive PC onto PCBus
IF	ϕ_1	Do tag compare; detect miss
	ϕ_2	Drive PC out to external cache
CM_1	ϕ_1	
	ϕ_2	Instruction back from external cache
CM_2	ϕ_1	
	ϕ_2	Write instruction into cache and instruction register
RF	ϕ_1	Miss sequence completed
	ϕ_2	

Figure 13: Miss Timing for a Buffer

miss penalty, it might increase the average cost of a fetch by increasing the cycle time of every instruction.

Clearly, if we require a minimum of two cycles to access the external cache and write the instruction into the instruction cache, the only way to reduce the miss penalty is to detect the miss sooner. If a miss can be detected before the end of ϕ_1 of IF, the PC can be driven out right away, eliminating one cache miss cycle. In the buffer scheme, driving the address to the cache array to fetch the valid bits causes the miss signal to appear late. The solution is to move the valid bits into the datapath along with the tags. Then, tag comparison and valid bit checking can be done a phase earlier.

There is one problem, however. To know which valid bit to fetch, we need to know which tag matched. Instead of doing the valid bit access after the tag comparison, we fetch all possible valid bits in parallel, one for each tag, along with the tag compare. The result of each tag compare is "AND-ed" with its corresponding valid bit. A cache miss occurs if none of these outputs is true.

Figure 13 shows the miss timing for a buffer with the valid bits and tags stored in the datapath. The miss penalty is now reduced to two cycles.

An interesting and important side effect of moving the tags and the valid bits into the datapath is that the cache array becomes strictly a RAM array. With the valid bits in the cache array, the array would need to be customized to our specific cache configuration because the valid bits must be cleared when a new tag is written into that block. Now, the valid bit circuitry is independent of the RAM and also simpler to implement.

With a two cycle penalty, and a 20% miss rate, the performance loss is 40% which is still significant. We can reduce the miss penalty to one cycle by combining *CM1* and *CM2* into one cycle. As mentioned before, this extends the machine cycle time. An alternate solution is to use the extra cycle wisely to prefetch another word.

4.5.3 Prefetching

A side effect of the two cycle miss penalty in MIPS-X is that the timing allows fetching back not only the instruction that missed but also the next instruction (not the next sequential one, but the instruction to be executed next) during the extra cache miss cycle. This means that the worst miss rate for the cache is 50%, and the average miss rate is about half of what it would be without fetching back two words on a miss.

The method of prefetching an extra word can be explained with the aid of the cache miss timing in Figure 14. In the phase following miss detection for the current instruction

IF_{-1}	ϕ_1	Compute PC_{miss}
	ϕ_2	Drive PC onto PCBus
IF	ϕ_1	Do tag compare and detect miss; Compute PC_{next}
	ϕ_2	Drive PC_{miss} out to external cache
CM_1	ϕ_1	
	ϕ_2	$Instruction_{miss}$ back from external cache
		Drive PC_{next} out to external cache
CM_2	ϕ_1	
	ϕ_2	Write $instruction_{miss}$ into cache and instruction register
		$Instruction_{next}$ back from external cache
RF	ϕ_1	Miss sequence completed
	ϕ_2	Write $instruction_{next}$ into cache and instruction register

Figure 14: Fetching Back Two Words on a Miss

(whose address is PC_{miss}), the address of the next instruction (PC_{next}) has already been calculated. This address would have been sent to the instruction cache in the normal sequence. While the external cache is being accessed, the next instruction address is set up on the address pads (in *CM1*); and while the missed instruction is written into the instruction cache during *CM2*, the external cache is accessed for the next instruction. Then, in the following phase (*RF*), when execution of the missed instruction is commenced, the next instruction is simultaneously written into the instruction cache and instruction register. Thus, the timing of the pipeline allows the prefetch to occur quite naturally.

Prefetching the extra word has a tremendous performance impact. For the MIPS-X hybrid buffer, the miss rate drops from 23.17% to 11.85%, or performance degradation drops from about 40% to 20%. Note that this scheme has the effective performance impact of a one cycle miss penalty, but without the risk of increasing the machine cycle time. Implementation is also simple because fetching the second word fits in with the natural flow of the cache miss sequence. This shows that careful matching of the cache miss timing to the pipeline of the machine can give significant performance benefits.

Other prefetching schemes that exploit the available excess bandwidth were also considered. For example, in MIPS-X, when the processor is not fetching data, the I/O pins are free and instructions could be prefetched into the cache without affecting any other activity of the processor. Load forward is another prefetch scheme where instructions are prefetched up to the end of the current cache block, a variant of which is used in the SPUR multiprocessor [7]. However, these schemes could not be used in MIPS-X because the instruction cache does not have sufficient bandwidth. MIPS-X uses 100% of the instruction cache bandwidth for fetches, preventing a prefetcher from using the cache. The instruction cache is only free during instruction cache misses. Thus, no prefetch scheme can do better without dramatically changing the cache organization.

4.6 Summary of Organization Choice

Table 2 summarizes the cache performance statistics for the various schemes assuming that two words are fetched back on a miss. Appendix A presents the individual benchmark miss rates for the four organizations in Table 3. Conventional cache organizations perform worse than buffers because of their high miss penalty. Note that the lowest miss rate does not yield the best performance. A hybrid buffer with four sets, associativity eight, and

block size 64 bytes performs best with an access time of 1.24 cycles, and is used in the MIPS-X design.

Cache type	Implement-able	Num. Sets	Set size	Block size	Miss rate(%)	Miss penalty	T_{ave}
C-cache	YES	256	1	4	16.74	3	1.50
C-cache	NO	16	8	16	10.90	3	1.33
Buffer	YES	1	8	256	14.78	2	1.30
Hybrid buffer	YES	4	8	64	11.85	2	1.24

Table 2: Summary of cache performance. Block size is in bytes; miss penalty and access time are in cycles.

5 Selection of Other Cache Parameters

5.1 Replacement

The replacement algorithm has traditionally been very important in cache design. Of the feasible schemes, least recently used replacement is considered to perform the best, although Smith and Goodman [22] show evidence that it might be inferior to random replacement for instruction cache design. After considering a number of replacement strategies we came to the conclusion that the replacement algorithm is not critical to the design when the cache size is small. We summarize our results in Figure 15 showing the performance of the MIPS-X hybrid buffer for five replacement schemes: LRU, RAND, FIFO, RING, and RING-M.

LRU (least recently used) is where the least recently accessed block in any given set is replaced. In random replacement (RAND), a truly random choice of block to be replaced is made. FIFO is first in first out, where the block present the longest in any given set is replaced first. RING, is a pseudo-random replacement scheme where a ring counter with the same number of states as the set size is maintained. The counter points to the block in each set that must be replaced on a block miss or if an address tag does not match any of the cache tags. The counter is bumped one state after every block miss. RING-M is a modification of the above scheme. The difference is that the ring counter is also bumped one state if the counter points to a block that matches the most recently addressed tag. This optimization reduces the probability of a recently used block being replaced.

Figure 15 compares the relative performance of our hybrid buffer for the various replacement schemes. RAND, FIFO, RING, and RING-M have about equal performance. LRU is slightly (10.91% vs. 11.85%) better than the other schemes. For the MIPS-X design, we chose one of the RING schemes because of its simpler implementation. The RING-M scheme was used to solve a subtle problem with the double fetch on instruction cache misses. As stated before, two instructions, I_{miss} and I_{next}, are fetched on a cache miss. The problem arises when the first word (I_{miss}) hits in the tags, but is not valid yet, and the next instruction (I_{next}) misses in the tags.[7] In the RING scheme, if the ring counter points to the block that I_{miss} will occupy, then that block will be replaced by the tag corresponding to I_{next}, causing I_{miss} to have nowhere to go when it is received from the external cache. To avoid this state, we bump the counter if it points to a cache block corresponding to the most recent address tag.

[7] I_{next} is not constrained to be in the same block as I_{miss}.

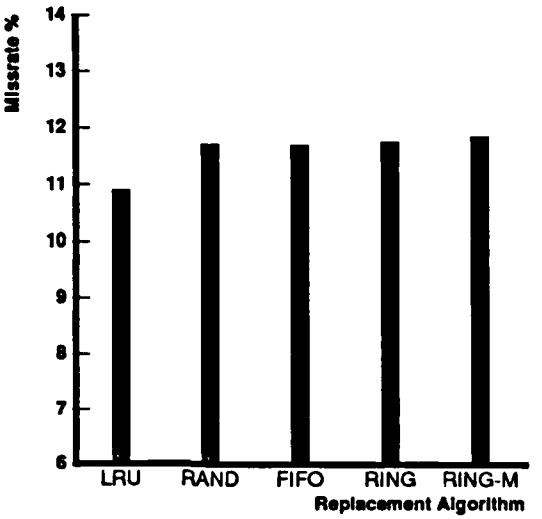


Figure 15: Comparison of replacement schemes.

5.2 Context Switch Mechanisms

Virtual caches, or caches addressed using the virtual address generated by the processor, have the advantage that virtual to physical translation is removed from the critical path. However, they have other multiprogramming related problems. For one, the integrity of a process address space is harder to maintain. A simple solution is to flush the cache on every process switch. The performance degradation due to cache flushes is not serious for small caches.

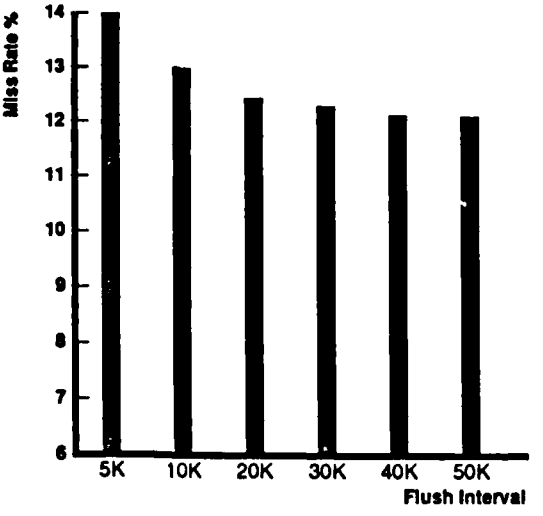


Figure 16: Effect of cache flushing on context switches.

Figure 16 shows the miss rate for a hybrid buffer flushed every Q instructions, where Q ranges from 5000 to 50000. A higher frequency of flushing is expected in time sharing workloads while batch jobs will be much lower.

Flushing the entire cache is easily achieved in VLSI by providing a cache reset signal. However, a cache flush would require a special instruction. To avoid defining another

instruction, we decided to use a simple software technique and trade off a little performance. The virtual address space is half system and half user. To flush the cache of user instructions, we cause the processor to jump to 32 specific system addresses making all the tags be in system space. This requires 32 extra instructions or 64 extra cycles[8] every cache flush. Even if cache flushing takes place, say, once every 20000 cycles, the miss rate would go up from 11.85% to 12.33%.

5.3 Testability Features

We have included a number of features into the design to enhance the testability of the instruction cache and the processor. The instruction cache and the rest of the processor have separate power supplies making it possible to power the processor independent of the cache. An external signal (ICacheDisable) forces the processor to always cache miss on instruction fetches allowing the processor to run even if the cache is not completely functional.

The size of the instruction cache forced us to include a method to directly access the RAM. When an external signal called ICacheTest is asserted, the PC is forced to generate sequential addresses. These addresses will initially miss in the cache, and data will be loaded from the data pads. After filling up the cache, the processor can be reset and the entire cache read. Although this interface does not allow us to perform random reads and writes into the cache, it does let us directly test the basic functionality of the cache before we use it for supplying instructions to the processor.

6 Conclusions

Cache design has already been studied in great detail but only recently has it been feasible to implement caches on the same chip as a processor. We have shown that for on-chip caches other considerations besides hit rate are important. These include the total usable area, the timing of cache accesses, the physical organization of the cache, and the aspect ratio of the resulting design. Minimizing the average instruction access time is the key goal. The average instruction access time is a combination of both the miss rate and the number of cycles needed to service a miss. We have shown that given several physical organizations that satisfy the space and size constraints, the resulting miss rate can vary considerably, and the organization with the lowest miss rate does not necessarily result in the best performance.

We have shown the importance of the tradeoffs between cache architecture and implementation by describing the design of a real on-chip cache for MIPS-X. Given an initial set of constraints for the physical dimensions and the cycle time of the desired cache, we used trace driven simulations to measure the performance of three basic cache configurations, varying several parameters such as associativity, set size, and block size. With these results, we computed the average instruction access times by taking into account the number of cycles needed to service a miss for each of the configurations, and made our choice based on the minimum average instruction access time. The result was a cache organization that is a hybrid between a conventional cache and an instruction buffer. This organization has a miss rate of roughly 12% for a set of large benchmarks, and results in an average instruction access time of 1.24 cycles. This penalty is roughly 3 times smaller

[8]Note that only 16 of the 32 instructions suffer cache misses.

than the penalty of our first cache organization, although the basic cache organization (cache size, block size, etc.) remained unchanged.

7 Acknowledgements

The MIPS-X research effort was supported by the Defense Advanced Project Research Agency under contract No. MDA903-83-C-0335. Paul Chow was partially supported by the Natural Sciences and Engineering Research Council of Canada.

We also gratefully acknowledge the contributions of Malcolm Wing, Karen Huyser, Scott McFarling, C. Y. Chu, Steve Richardson, Steve Tjiang, Richard Simoni, Don Stark, Glenn Gulak, and Steven Przybylski.

A Individual benchmark cache miss rates

Table 3 presents the individual miss rates for all the benchmarks. Miss rates for Dnf, Gc, and Opt, are lower than average because in Dnf, the program kernel is small; Gc spends about half its time in the garbage collector which is a small body of code; and Opt is very iterative.

		Miss rates %			
Benchmark	Instr. refs.	C-cache (256,1,4)	C-cache (16,8,16)	Buffer (1,8,256)	H. buffer (4,8,64)
Bigfm	3042360	14.47	9.03	14.90	10.00
Dnf	1150743	15.57	2.52	23.23	3.96
Hopt	2595403	18.12	10.93	15.38	11.33
Simu	3095037	16.90	12.90	16.47	12.42
Upas	1680149	20.05	15.62	26.45	17.91
Comp	230240	20.45	13.98	22.17	16.48
Frl	1221133	27.13	20.18	28.11	23.28
Gc	1324951	9.46	8.10	8.74	5.48
Opt	3055475	12.53	6.80	16.71	8.64
Rat	673256	28.12	18.93	35.73	22.71
Average		16.74	10.90	18.46	11.85

Table 3: Summary of cache miss rates for individual benchmarks. Cache type is specified as a triple (S,D,B), where S is number of sets, D is degree of associativity, and B is block size in bytes.

References

[1] J. L. Hennessy. VLSI Processor Architecture. *IEEE Transactions on Computers*, c-33(12), December 1984.

[2] R. L. Sites. How to Use 1000 Registers. In *Proceedings, 1st Caltech Conf. VLSI*, January 1979. California Institute of technology, Pasadena, CA.

[3] James R. Goodman. Using Cache Memory to Reduce Processor-Memory Traffic. In *Proceedings of the 10th Annual Symposium on Computer Architecture*, pages 124–131, June 1983.

[4] D. A. Patterson and C. H. Sequin. Design Considerations for Single-Chip Computers of the Future. *IEEE Transactions on Computers*, C-29(2):108–116, February 1980.

[5] G. Radin. The 801 Minicomputer. In *Proc. SIGARCH/SIGPLAN Symp. Architectural Support for Programming Languages and Operating Systems*, pages 39–47, March 1982. ACM, Palo Alto, CA.

[6] M. Horowitz and P. Chow. The MIPS-X Microprocessor. In *Proc. WESCON 85*, 1985. IEEE, San Francisco.

[7] M. D. Hill, S. J. Eggers, J. R. Larus, G. S. Taylor, G. Adams, B. K. Bose, G. A. Gibson, P. M. Hansen, J. Keller, S. I. Kong, C. G. Lee, D. Lee, J. M. Pendleton, S. A. Ritchie, D. A. Wood, B. G. Zorn, P. N. Hilfinger, D. A. Hodges, R. H. Katz, J. Ousterhout, and D. A. Patterson, *SPUR: A VLSI Multiprocessor Workstation.* Computer Science Division 86-273, University of California, Berkeley, December 1985.

[8] Doug MacGregor, Dave Mothersole, and Bill Moyer. The Motorola MC68020. *IEEE Micro*, 101–118, August 1984.

[9] First Look at Motorola's Latest 32-Bit Processor. Electronics, September 18, 1986.

[10] George S. Taylor, Paul N. Hilfinger, James R. Larus, David A. Patterson, and Benjamin G. Zorn. Evaluation of the SPUR Lisp Architecture. In *Proceedings of the 13th Annual Symposium on Computer Architecture*, pages 444–452, June 1986.

[11] Alan Jay Smith. Cache Memories. *ACM Computing Surveys*, 14(3):473–530, September 1982.

[12] J. S. Liptay. Structural Aspects of the System/360 Model 85, Part II: The Cache. *IBM Systems Journal*, 7(1):15–21, 1968.

[13] Alan Jay Smith. Cache Evaluation and the Impact of Workload Choice. In *Proceedings of the 12th Annual Symposium on Computer Architecture*, pages 64–73, June 1985.

[14] Anant Agarwal, Richard L. Sites, and Mark Horowitz. ATUM: A New Technique for Capturing Address Traces Using Microcode. In *Proceedings of the 13th Annual Symposium on Computer Architecture*, pages 119–127, June 1986.

[15] J. L. Hennessy, N. Jouppi, S. Przybylski, C. Rowen, and T. Gross. Design of a High Performance VLSI Processor. In *Proceedings, Third Caltech Conf. VLSI*, pages 33–54, March 1983. California Institute of technology, Pasadena, CA.

[16] K. R. Kaplan and R. O. Winder. Cache-Based Computer Systems. *Computer*, 6(3):30–36, March 1973.

[17] W. D. Strecker. Cache Memories for PDP-11 Family of Computers. In *Proceedings of the 3rd Annual Symposium on Computer Architecture*, pages 155–158, January 1976.

[18] Donald Alpert. *Performance Tradeoffs for Microprocessor Cache Memories* . Computer Systems Laboratory 83-239, Stanford University, December 1983.

[19] *CRAY-1 Computer Systems, Hardware Reference Manual.* Cray Research, Inc., Chippewa Falls, WI, 1979.

[20] *VAX-11 Architecture Reference Manual.* Digital Equipment Corporation, Bedford, MA, 1982. Form EK-VARAR-RM-001.

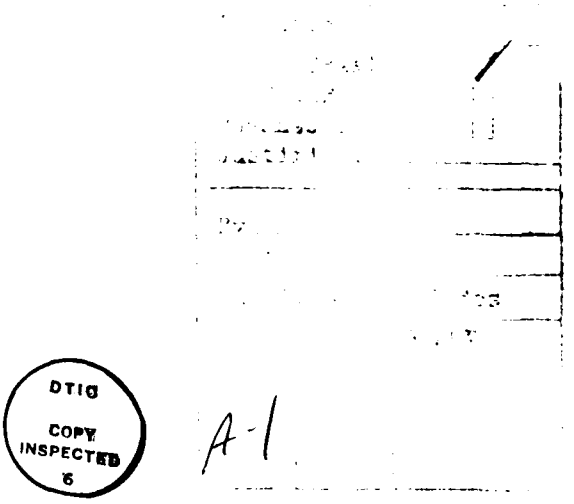
[21] Mark Hill and Alan Jay Smith. Experimental Evaluation of On-Chip Microprocessor Cache Memories. In *Proceedings of the 11th Annual Symposium on Computer Architecture*, pages 158–166, June 1984.

[22] James E. Smith and James R. Goodman. A Study of Instruction Cache Organizations and Replacement Policies . In *Proceedings of the 10th Annual Symposium on Computer Architecture*, pages 132–137, June 1983.

[23] Anant Agarwal, Mark Horowitz, and John Hennessy. *An Analytical Cache Model.* Computer Systems Laboratory 86-304, Stanford University, September 1986.

[24] D. Alpert, D. Carberry, M. Yamamura, Y. Chow, and P. Mak. 32-bit Processor Chip Integrates Major System Functions. *Electronics*, 56(14):113–119, July 1983.

[25] Paul Chow. *MIPS-X Instruction Set and Programmer's Manual.* Computer Systems Laboratory 86-289, Stanford University, May 1986.



END

3-87

DTIC